## Sufficient Statistics

When learning MLE, we find that usually the likelihood function (i.e., the joint pdf/pmf) can be factorized into two parts

$$f(x_1, \ldots, x_n; \theta) = K_1\big[u(x_1, \ldots, x_n); \theta\big] \cdot K_2(x_1, \ldots, x_n),$$

where $K_2(\cdot)$ does not depend on $\theta$, and $u(x_1, \ldots, x_n)$ is a summary statistic (e.g., sample mean/variance) of the data. We call $u(X_1, \ldots, X_n)$ a **sufficient statistic** for $\theta$.

An equivalent definition is that if

$$\frac{p(x_1, \ldots, x_n \mid \theta)}{p(u(x_1, \ldots, x_n) \mid \theta)} \tag{1}$$

does not depend on $\theta$ then $u(X_1, \ldots, X_n)$ is a sufficient statistic for $\theta$, where $p(u \mid \theta)$ denotes the pdf/pmf of $u(X_1, \ldots, X_n)$. That is, $u(X_1, \ldots, X_n)$ has exhausted all the relevant information about the unknown parameter $\theta$ contained in this random sample.

Go through examples in `SuffStatans.pdf`.

Sufficient statistics are not unique, and not necessarily of the same dimension.

- if $u(x_1, \ldots, x_n)$ is a sufficient statistic, then any one-to-one transformation of $u$, say $v(x_1, \ldots, x_n)$ would also be a sufficient statistic. For example, if $\sum_i X_i$ is sufficient, so are $\bar{X}$ and $1/\bar{X}$.
- $(X_{(1)}, \ldots, X_{(n)})$, an ordered version of the sample $(X_1, \ldots, X_n)$, is a sufficient statistic, which doesn't make that much of data reduction, since it's of the same dimension as the original data.
- If $\bar{X}$ is sufficient, so is $(\bar{X}, X_1)$: if $\bar{X}$ has extracted all the relevant information in the sample related to $\theta$, $(\bar{X}, X_1)$ wouldn't contain any less, so intuitively it must be sufficient too.

## Minimal Sufficient Statistics

From the discussion above, we see the need of getting the most compact sufficient statistic. A statistic $u^*(X_{1:n})$ is **minimal sufficient** if

1. it is sufficient and
2. for any other sufficient statistic $u(X_{1:n})$, $u^*(X_{1:n})$ is a function of $u(X_{1:n})$.

In other words, a minimal sufficient statistic achieves the maximal data reduction without losing any information about $\theta$.

Minimal sufficient statistic isn't unique since any one-to-one transformation of $u^*$ would still be minimal sufficient.

In most examples in `SuffStatans.pdf`, the derived sufficient statistics are minimal sufficient, which are often of the same dimension as the unknown parameters. For example, in Example 4, for $\mathsf{N}(\mu, \sigma^2)$ where both parameters are unknown, the derived sufficient statistic is of two-dimension: the sample mean and sample variance. But there are some exceptions:

- $X_1, \ldots, X_n$ iid $\sim \mathsf{DE}(\theta)$ where "DE" stands for Double Exponential. The corresponding pdf is given by

  $$f(x;\theta) = \frac{1}{2} e^{-|x-\theta|}, \quad x \in \mathbb{R}.$$

  $$\prod_{i=1}^{n} f(x_i;\theta) = \frac{1}{2^n} \exp\Big\{ -\sum_i |x_i - \theta| \Big\},$$

  which cannot be further simplified, so the minimal sufficient statistic is just the data $(x_1, \ldots, x_n)$ (or its ordered version) itself.

- $X_1, \ldots, X_n$ iid $\sim \mathsf{Unif}(\theta, \theta+1)$. In class we have shown that $(X_{(1)}, X_{(n)})$ is a sufficient statistic for $\theta$. In fact, no more reduction we can apply on the data, that is, $(X_{(1)}, X_{(n)})$ is minimal sufficient. Here the minimal statistic is not of the same dimension as the unknown parameter $\theta$.

  Define $R = X_{(n)} - X_{(1)}$ and $M = (X_{(1)} + X_{(n)})/2$. Then $(R, M)$ should also be a minimal sufficient statistic for $\theta$ (since the transformation is one-to-one).

  Note that

  $$R = X_{(n)} - X_{(1)} = (X_{(n)} - \theta) - (X_{(1)} - \theta) = Z_{(n)} - Z_{(1)},$$

  where $Z_{(n)}$ and $Z_{(1)}$ are the max and min for random samples $Z_1, \ldots, Z_n$ iid $\sim \mathsf{Unif}(0, 1)$. That is, the distribution of $R$ does not depend on $\theta$.

  Naturally, one would ask whether we could drop $R$ out of the minimal sufficient statistic and only keep $M$ since the distribution of $R$ has nothing to do with $\theta$.

  In this case, although $R$ alone does not provide any information about $\theta$, it does when being combined with other statistics: if $R$ is very large, say, close to 1, then we would know that $X_{(1)}$ and $X_{(n)} - 1$ are very close $\theta$, so $R$ provides information on the precision of the estimate of $\theta$.